**ABSTRACT:**

Soybeans are a vital global crop rich in protein and oil, making them essential for food production, livestock feed, and biofuel. Now a days it is impossible and time consuming for humans to predict the defects in soybean seeds. Since soybeans are useful in today global market it is essential for identifying the defects in soybean seeds to improve the production and also the quality of soybeans. Our objective is to predict the defects in soybean seeds using machine learning models. Initially the soybean seed data set was extracted from nearly 5000 different kinds of soybean seed images. The second phase is to develop a suitable machine learning model that can predict the defect of soybean seed. For easiness we had categorized the soybean seed into: complete, broken, immature, spotted, and skin-damaged. The dataset includes key seed features such as size, weight, color, shape and using these key features and other supporting features we had predicted the basic defects in soybean seeds.

**INTRODUCTION:**

Soybean seed classification, a very important tool in agriculture, has been pivotal in crop quality and market value in the soybean business. Soybeans are used in food production, animal feed, and other biofuels. Traditionally, seed companies offered a range of seed options to farmers based on multi-year yield performance trials conducted at various test farm locations. Farmers then choose seed varieties that are suitable for their farm's soil and climatic conditions, optimizing their crop yield. Quality soybean seeds, depending on characteristics such as size, shape, color, and physical damage, affect crop productivity, nutrition, and processing efficiency.While traditional manual inspection has been the norm for evaluation of seed quality, it is time consuming and prone to human error. There is thus a great need for machine learning as a transformative solution in the automation of and the accuracy of seed classification. In this project, we use ML algorithms in the classification of soybean seeds, based on numerical data obtained from physical characteristics that are elicited by images from the seeds.Image-based classification techniques, although effective, heavily rely on computational resources for their processing; our approach uses structured datasets in CSV format with encapsulated numerical attributes like seed weight, dimensions, moisture content, and others. The datasets are further enhanced with features extracted from seed images, including edge detection and shape analysis, texture metrics, color distribution, and correlation measures. These features play a critical role in seed physical condition assessment and defect identification.

Key features such as edge and shape analysis capture geometric properties like seed boundaries and contour irregularities, enabling the identification of broken or misshaped seeds. Texture and correlation metrics, derived from the Gray-Level Co-occurrence Matrix (GLCM), help detect visual defects such as discoloration or contamination. Color distribution provides insights into seed maturity and potential damage due to environmental factors, while Histogram of Oriented Gradients (HOG) features highlight subtle surface irregularities and distortions.Preprocessing involved applying necessary scaling, normalization, and dimensionality reduction for optimizing model performance. However, these techniques help the models to increase their efficiency as training is quicker and the performance is not hampered. We used Random Forest, SVM, and Gradient Boosting algorithms that are suitable for working on numerical data because they can easily handle diverse feature sets.A hybrid stacking classifier combining XGBClassifier and Logistic Regression attained the best performance for our dataset, at 82% accuracy, with precision of 81%, recall of 82%, and F1-score of 82%. This approach shows that machine learning is a suitable solution in the automation of soybean seed classification. In this way, it could easily improve speed, scalability, and accuracy beyond traditional methods.

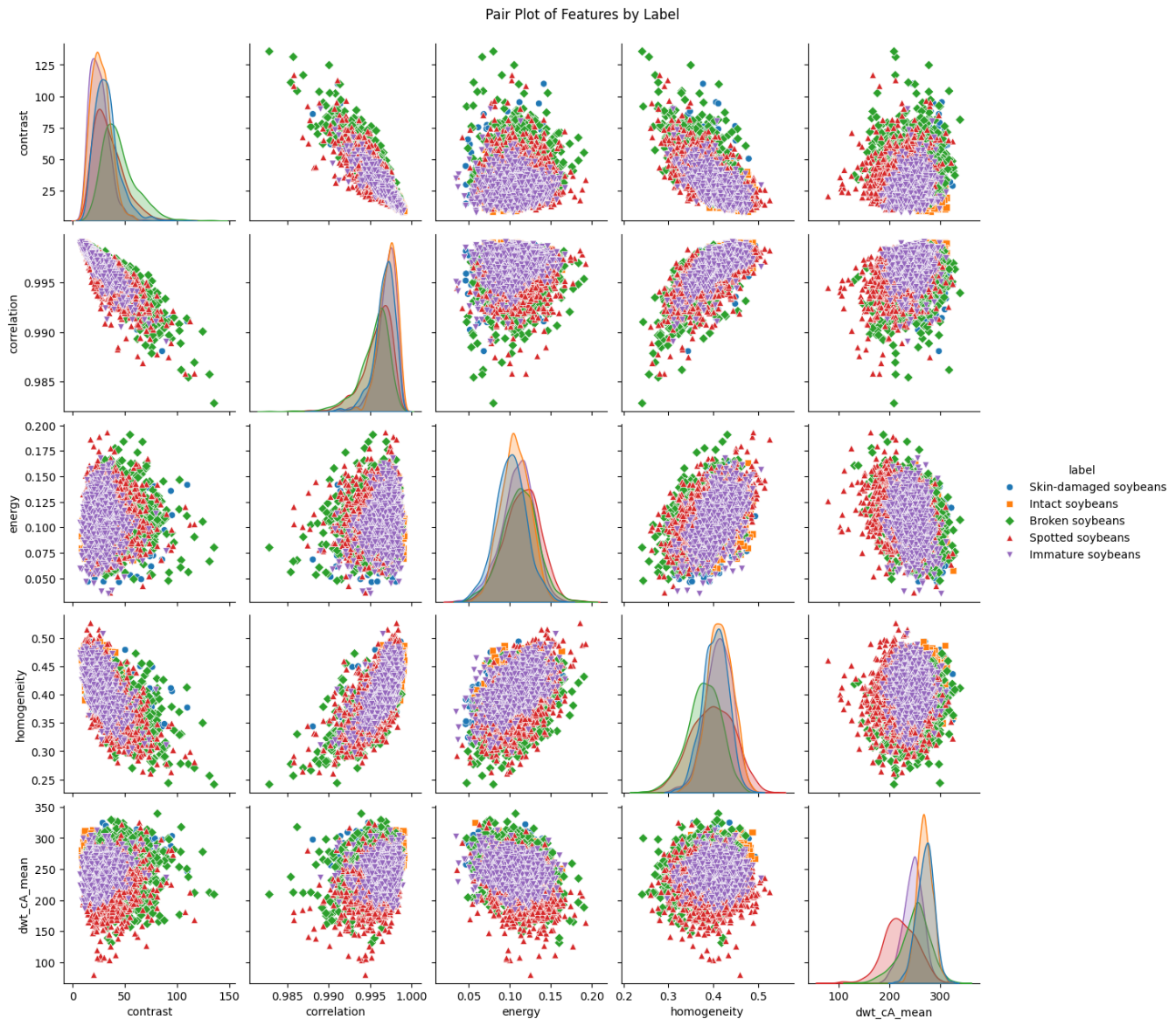
**LITERATURE SURVEY:**

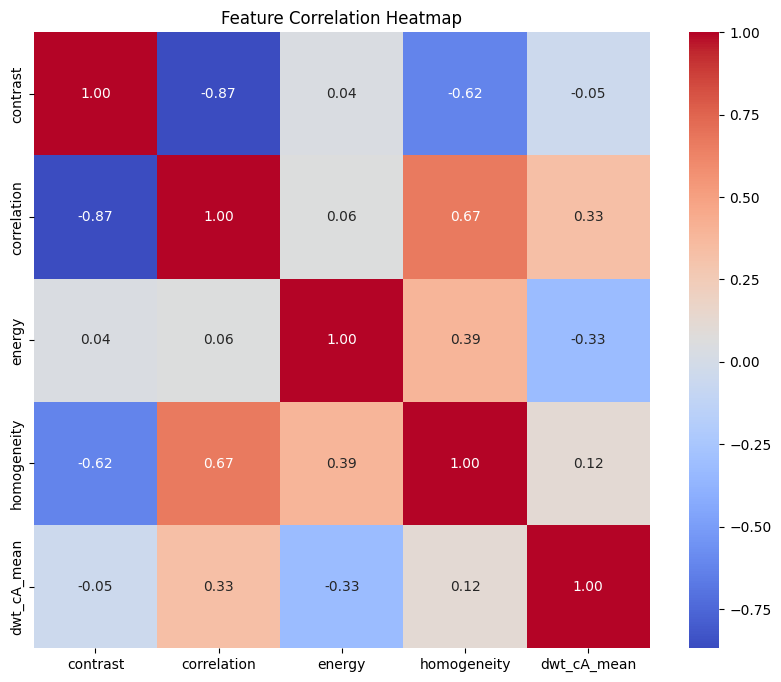
* The classification of soybean seeds is a vital but very time-consuming process for Brazilian agribusiness cooperatives. It often poses logistical problems, like waiting longer and queuing, which can severely affect the cost side of operations. This study proposes an innovative approach for the estimation of average lead times and queue durations in the soybean seed classification process to make decisions about manpower and machinery deployment.The proposed methodology integrates Discrete Event Simulation with Machine Learning algorithms to provide a robust framework for optimizing logistical efficiency. The approach is structured in three key phases and offers flexibility to deal with various logistics-related decision-making challenges.Conducting on-site assessments to gather very detailed, process-specific data essential for the right modeling. Design a discrete event simulation model that simulates this classification process to generate synthetic datasets capturing the dynamics of the system under different conditions.The prediction phase uses the machine learning algorithms to analyze the simulated data and predict lead times, as well as queue durations.This enables a more integrated approach by improving predictive precision and strategic planning. It arms agribusiness cooperatives with actionable insights for improvement in soybean seed classification processes, thus improving cost management and business performance in logistics. **[1]**
* In this paper, a computational framework to detect and quantify defects in soybean seeds through the utilization of advanced deep learning techniques is provided. Toward the effective differentiation between healthy and defective seeds, the authors propose a novel lightweight model, the Soybean Seed Defect Identification Network (SSDINet).The methodology initiates with the development and preprocessing of a soybean seed dataset by an in-house-designed seed contour detection algorithm for extracting features from seeds. Classification architecture integrates a Convolutional Neural Network, depthwise convolution blocks, and squeeze-and-excitation blocks, which end up generating a computationally efficient, faster, and more accurate model than current approaches.Performance evaluation has shown that SSDINet performs better than the current state-of-the-art models, showing its accuracy in detecting soybean seed defects. The combination of lightweight design and high accuracy makes SSDINet a robust tool for automated quality assessment in soybean seed classification. **[2]**
* Seed germination and vigor determination are essential in the evaluation of the performance of various seed lots, thus improving the storage and sowing processes efficiency. However, the complexity and large amount of data generated through various tests used to determine the seed quality leave much to be desired for manual quality control.This study aims to address this challenge by leveraging machine learning (ML) techniques to evaluate the physiological quality of soybean seeds and rank seed lots based on their quality. The evaluation was conducted in two phases.The first immediately after harvest and the second following six months of storage.Several ML models were used, including Random Forest, Multi-Layer Perceptron (MLP), J48, and Classification via Regression. Among the these models, Random Forest and Classification via Regression were most accurate, thus being the best seed lot ranking models. This approach thus offers a systematic and efficient method in soybean seed quality assessment, thus offering valuable insights that would optimize seed storage and usage strategies. **[3]**
* Sorting high-quality soybean seeds represents an important and labor-intensive activity within the process of quality assurance and food safety. This contribution proposes a full pipeline for soybean seed classification using segmentation-classification workflow for process simplification.Towards the classification step, a new lightweight model is presented, known as Soybean Network (SNet). Based on convolutional neural networks, SNet employs Mixed Feature Recalibration modules to further enhance the network's ability in focusing on important features as well as ensure computational efficiency.The proposed SNet has demonstrated a high accuracy that surpassed the performance of six already existing state-of-the-art models. Its lightweight architecture really makes it especially effective in automatic soybean seed recognition on resource-constrained platforms, because it provides a practical solution toward efficient and reliable seed quality inspection. **[4]**
* Soybean (Glycine max) is a crucial global crop, rich in proteins and essential nutrients, vital for food security.High seed quality is necessary for optimal yields, but environmental and post-harvest factors can cause damage and degrade seed performance.Traditional methods for assessing seed quality, like visual inspection and chemical tests, are often destructive and time-consuming.Machine learning, particularly interactive Machine Learning (IML), offers a more efficient solution by combining human expertise with algorithms to improve accuracy, especially for complex datasets. Tools like llastik, which are user-friendly and open-source, have shown promise in other research fields but are not traditional machine learning to classify soybean seeds and seedlings, linking their appearance with physiological performance to enhance seed quality assessment.**[5]**
* The current paper focuses on the high-oil-content soybean seed identification with histogram-based imaging and a one-dimensional convolutional neural network model. A total of 5,510 samples distributed in 58 soybean varieties were analyzed across a certain spectral range. Using the 1D-CNN model, superior results in terms of accuracy were achieved compared to classical machine learning models that included Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Partial Least Squares Discriminant Analysis (PLS-DA). Among the four preprocessing techniques applied, Multivariate Scattering Correction (MSC) yielded the most effective results.In the KNN approach, the algorithm calculates the distance between feature values of test and training data points, selects the k nearest categories based on minimal distance and classifies based on majority voting. Further, the Successive Projections Algorithm (SPA), a commonly adapted forward selection algorithm for spectral feature wavelengths, was used for feature selection refinement. **[6]**
* Soybean farming is a highly significant production in the world. The aim of this work is to develop climate-based yield prediction models by incorporating machine learning techniques for enhanced agricultural productivity. The framework of the research is divided into the following two stages:Calibration Stage: Meteorological data and soybean yield records are utilized from 47 locations of Mato Grosso do Sul to train and calibrate ML algorithms.Prediction Stage: The best-performing algorithm was used in estimating soybean yields for all the state areas.
* Multiple Linear Regression, Multilayer Perceptron, Support Vector Machine, Random Forest, and Extreme Gradient Boosting were used in the machine learning methods. Data splitting was applied with 30% for validation and 70% for training to warrant good performance on the models.The calibrated models were quite precise and accurate during the training and validation phases. Crucial climate variables highly correlated with yields included 26\_12\_ARM, which is positively correlated and related to the grain-filling stage, and 2\_10\_TDEW, a variable that is negatively correlated and related to emergence. This shows that climatic factors play an essential role in certain phenological stages of soybean yields.**[7]**
* This research proposes a supervised machine learning methodology for soybean classification, which will be presented as a technology to classify whether the soybean classification is standard or not. Widely applied in agriculture include weather consultation tools, agricultural commodity price predication systems, property management applications, as well as precision machinery, says Ferraz & Pinto (2017). Those tools offer decisive powers through the reduction of risk, maximization of profit, and allocation of the best resources.Emerging technologies such as artificial intelligence, data science, big data, the Internet of Things (IoT), and machine learning all contribute to the technological revolution in agriculture. These innovations back up different parts of the production chain to develop profitability, efficiency, and yield. Supervised machine learning has been very well developed and also integrated into the agricultural domain, delivering highly accurate and reliable results.This research presents a new supervised machine learning approach to predict the standard and non-standard classification of soybean grains. It incorporates the design of a website that adopts the intelligent computing technique in generating real-time predictions strictly aligned with MAPA normative instruction. Such an innovative approach will expedite an efficient, robust, and accurate soybean grain classification into standard and non-standard classification thus assertively informing decision-making processes and can push forth the adoption of technology in agriculture. **[8]**
* Weed management is one of the most critical components in soybean production and requires effective discriminating tools between crops from weeds. This research investigates the incorporation of NDVIs obtained from multispectral leaf reflectance data along with the Random Forest, a type of machine learning model, to detect soybeans from three broadleaf weeds, specifically Palmer amaranth, redroot pigweed, and velvetleaf.Images were taken on two controlled greenhouse experiments, from which twelve NDVIs were calculated: advanced vegetation indices, shortwave infrared water stress indices, and pigment indices. Then the Random Forest algorithm was used to classify plant types, successfully distinguishing between the soybeans, velvetleaf, and the two pigweed species. Classification performance improved when the two species of pigweed were classified together in a class.The results demonstrate that NDVIs were equivalent, or even superior to directly using multispectral bands for classification. The study emphasizes the fact that the success of remote sensing in weed detection is highly dependent on the spectral sensitivity of the devices and the efficacy of the algorithms used. This approach presents a promising direction to further precision agriculture by accurate discrimination between crops and weeds. **[9]**
* This study focuses on developing classification models for differentiating soybean seeds on the basis of physical features such as shape, size, and mass. Several machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree, were used to assess classification performance.Results suggest that Multilayer Perceptron (MLP) and Random Forest (RF) seem to be excellent classifying models for soybean varieties. RF, on the other hand, constructed from multiple decision trees, classifies objects by having each tree vote for a class. Finally, with majority voting, the final classification is taken robustly and accurately.The Pillai Trace and Wilks' Lambda showed that the differences of the physical traits of soybean varieties were statistically significant. This paper is a contribution to existing literature because it explains the classification of soybean seeds into binary systems, which had already been researched but are now improved with these results. This approach showcases the strength of machine learning in soybean seed classification and quality analysis. **[10]**
* Each year, agribusinesses, through their research and development efforts, introduce new soybean seed varieties with optimized traits for specific planting environments. For the purpose of guiding the farmer towards the most suitable varieties available in this competitive marketplace, the current study develops a decision support tool using data from Syngenta, an agribusiness company. The tool employs machine learning to predict yield performance across the range of weather and soil conditions under which soybeans might be planted.The yield data of different varieties of soybean were assessed using this dataset across multiple farms in the Midwest United States. Soybean farmers make critical decisions every year about the mix of varieties to plant. This tool will enhance decision making through data-driven insight, enabling optimization in crop selection for the highest yields.The impact of this tool will be huge; crop yields increase and consequently farm revenue, facilitated by evidence-based tailored decision making. This approach can also be scaled up and applied to other crops, contributing to the global initiatives towards addressing food security challenges by boosting agricultural productivity.**[11]**
* This research aims to develop a supervised machine learning-based approach for soybean classification and present it as an innovation in technology that determines whether the classification of a soybean is standard or not. According to Ferraz & Pinto (2017), some technologies employed in agriculture include weather consultation tools, agricultural commodity price prediction systems, property management applications, and precision machinery, among others. These tools make decision-making better since they tend to reduce risks, maximize profit, and optimize resource use.Such emerging technologies as artificial intelligence, data science, big data, the Internet of Things (IoT), and machine learning are technological leaders that govern this technological revolution in agriculture. These innovations support several components of production chains and lead to profit maximization, efficiency, and yield. Supervised machine learning has been highly developed and integrated into the agricultural domain, where it gives very accurate and reliable results.This research proposes a new machine learning supervised approach to the prediction of standard and non-standard classification for soybean grains. This involves developing a website using intelligent computing techniques that will present predictions in real-time following the normative instruction by the MAPA. This is a new and more efficient robust method for classifying soybean grains into accuracy, thus enabling broad decision-making action about it by the farming fraternity and promoting technological uptake in agriculture. **[12]**
* A modified Inception V3 model was applied in this study to classify defective soybean seeds, utilizing a dataset of 5,513 images categorized into five different classes. The F1-score, balancing precision and recall, showed outstanding performance across all classes. Advanced techniques in transfer learning, adaptive learning rate adjustment, and model checkpointing were applied to enhance its accuracy. Its precision and recall rates were very high at the end, emphasizing the usefulness of the model.This automated solution addresses the traditionally manual and time-intensive process of seed quality assessment, providing a more efficient and scalable alternative. The results resonate with the transformative potential of deep learning for modernizing agriculture practices. Overall, this study is a major breakthrough in agricultural technology, with an innovative and effective methodology for seed classification.**[13]**
* The web application of a classifier for soybean seeds based on deep learning increases the effectiveness of soybean quality assessment. This approach preprocesses seed images captured under non-uniform illumination, where the seeds are classified as normal, damaged, abnormal, or not classifiable. Classification in this work was implemented through lightweight CNNs, thus providing a critical enhancement in comparison with the commonly used traditional methods, which are mainly based on human observation and often inefficient and subjective.CNNs showed great promise in the non-destructive analysis of agricultural products. The model was deployed on the NVIDIA Jetson TX2, which achieved fast processing times per seed, fulfilling real-time evaluation requirements. The results showed that it is highly effective for real-time soybean seed quality assessment, especially in resource-scarce conditions, making it a practical and efficient solution for modern agricultural practices.**[14]**
* Seed quality stands as an important factor that directly influences seedling performance in the field. It is closely related to the stability of the genotype of the seed, and thus, genome integrity is essential for optimal growth. In order to determine this, seed germination and vigor tests were performed, while genomic DNA was extracted from the seed root meristems for analysis using electrophoresis and the comet assay to detect DNA damage.DNA integrity varied significantly as evidenced from the results, establishing a relationship between seed vigor and genome stability. Such findings point towards the requirement of maintaining DNA stability for ensuring quality seeds with further implications in crop productivity. Molecular markers could be devised from such studies to monitor and assess seed quality in future, thereby further advancing the generation and application of improved seeds and agricultural practices..**[15]**
* This study explains Genomic DNA which was extracted from root meristem of seeds from a lot of different resources available.The physiological characteristics of germination and vigor makeup seed quality. A complicated characteristic, seed vigor includes stress tolerance in the early phases of seedling development, seed longevity, and seedling growth rate. Preventing damage to seeds and improving their quality are major areas of focus. Thus, it is critical to identify low-quality lots and dispose of them in order to avoid subpar field performance.Genotoxicity in soybean seeds is linked to cellular damage in seeds of various genotypes, indicating that continuous cell essential functions that support embryo growth and healthy seedling emergence depend on intact DNA.This study set out to assess the physiological potential of soybean seeds (cv. BMX Potência RR) from various lots as well as look into the genomic DNA of the seeds' electrophoretic profile and DNA integrity in vivo (using the Comet assay).It displays the biometric information, which includes the seedling length and fresh and dry weight. The results show that the shoot length (SL), shoot fresh mass (SFM), root fresh mass (RFM), and root dry mass (RDM) of the soybean seed lots varied significantly (p < 0.05). Nevertheless, there were no discernible variations in vigor across the seed lots according to the findings of the assessments of root length (RL) and shoot dry mass (SDM) .**[16]**
* This study investigated the mutation of two QTL genes, GmHs1-1 and GmQHS1, and their role in regulating seed coat hardness in soybeans. Our focus was on the seed coat permeability phenotype of legume species, and two notable publications from 2015 drew our attention due to their significant findings regarding soybeans. Recent study by Sun et al. [9] showed the major quantitative trait locus (QTL) gene GmHs1-1, which encodes a calcineurin-like protein, to be a significant contributor to soybean seed coat impermeability. qHS1, another important QTL gene, was also discovered; it affects seed coat impermeability in soybeans. Two genes are found close to each other in the physical map: a suggestion of their interaction in the improvement of seed coat permeability in cultivated soybeans.Actually, although altering this trait might affect the life span of the seeds under specific conditions, the researchers concluded that changing these genes is vital for improving seed coat permeability: this would favor better seed germination and utilization in agriculture. This research advances our knowledge of the genetics mechanisms causing seed dormancy and what lessons could be taken from the manipulation of these genes to increase crop yields.**[17]**
* The genetic factors, which play a crucial role in determining soybean seed weight and size, are the main focus of the study. Seed weight and size are two major traits that influence crop production significantly. Understanding genetic mechanisms underlying seed size and weight is crucial. Crossing large-seeded soybean variety "Kebaliang" and small-seeded variety "SUZUMARU" resulted in segregating populations. The hundred-seed weight, width, thickness, and length of the seeds were measured over two generations.This work greatly enhances the knowledge of the genetics controlling these traits. As a first step, this study used bulk segregant analysis (BSA) to rapidly pinpoint putative genetic markers of interest associated with the desired trait. Thirty quantitative trait loci (QTLs) for seed weight and size were tagged, as well as potential genes within these QTLs that proved stable and their expression patterns and sequencing variations.**[18]**
* This study presents a novel approach by combining reinforcement learning and Support Vector Machine (SVM) models with spatial frequency domain imaging in the identification of insect-damaged soybean seeds. Soybean seeds tend to suffer a direct reduction in quality, especially in case of attack by pests like Riptortus pedestris. The conventional methods for inspecting the soybean seeds are time-consuming, making machine learning and advanced imaging techniques useful for automating it.The introduced method employs optical coefficients for the purpose of detecting subsurface hidden damages that are unable to be observed by a naked eye through traditional methods. The preface underlines the significance of soybeans as a high-nutrition crop and highlights the intractability that pest-related damage poses. Manual inspection and conventional machine vision, typically performed methods, do not suit the early signs of damages.Reinforcement learning has been added to increase the accuracy and recall of many models including SVM, GRNN, and MLR while showing noticeable improvement with the use of the RL-SVM model. Experimental results prove that model performance is strongly increased with the incorporation of reinforcement learning, and the achieved Macro-Recall in the case of the RL-SVM is 0.9635. This methodology ensures higher accuracy in detecting pest damage, which adds to the quality evaluation of seed and increases crop yield.**[19]**
* This paper explores identification of a major quantitative trait locus, GmSW17, which has a critical influence on the regulation of soybean seed size. It encodes for the ortholog of the ubiquitin-specific protease family, specifically UBP22, which is the member of the USP/UBP family. Recent studies have revealed functional properties of several ubiquitin-related genes involved in seed size determination. The estimated global growth rate of population and rising living standards means that the soybean yield needs to multiply by at least two fold by 2050. Several factors control the traits of soybean production, and seed size is an important one for high-yield soybean breeding. High-yield soybean varieties can only be bred by identifying major genes governing seed size and understanding their regulatory mechanisms. The study concludes that though not fully stabilised, the two haplotypes of GmSW17 have successfully been integrated into soybean breeding programs from various regions.**[20]**
* This paper discusses the development of the DSSAT-CROPGRO-Soybean model that proved effective in simulating key soybean seed traits: emergence, flowering, and maturity. Theoretical basis to the simulation results provides a new framework in soybean production in southern Xinjiang. A good irrigation system sets out to be one of the most effective strategies to ensure consistent high crop yields and at the same time conserve water. Soybean is particularly sensitive to excess water application, hence the need for an appropriate irrigation system.The study indicated that different irrigation levels greatly affected the soybean yields. Two years of field trial data (2022 and 2023) were employed for calibrating and validating the DSSAT-CROPGRO-Soybean model. Meteorological data spanning from 1994 to 2023 was also integrated into the simulation to generate potential soybean yield and biomass. The research analyzed sustainability and stability in soybean yields under various irrigation regimes for drip-irrigated soybean under mulch in southern Xinjiang. It mainly aimed at finding an efficient technique for irrigation that could make soybean farming more lucrative as well as sustainable.Ultimately, the conclusion drawn from the study was that DSSAT-CROPGRO-Soybean is a very effective model for predicting phenological stages, biomass, and yield in soybeans. This means the model can be used to optimize the soybean irrigation system in southern Xinjiang to better manage water and increase crop yields.**[21]**
* The study investigates the use of laser-induced breakdown spectroscopy in combination with deep learning techniques for rapid identification of various soybean seed varieties, an important component for agricultural standardization and food processing. Thus, with the overall objective of research development, researchers developed a time-efficient sample preparation technique by pressing soybean seeds into rubber sand-filled culture plates instead of grinding or squashing. Spectral data were obtained from each seed; from this, a majority vote system was applied to classify the seed variety based on three spectra per seed. The experiment made comparisons over several machine learning and deep learning models, particularly in regards to CNN architectures. For visualization of the seed variety clusters, t-distributed stochastic neighbor embedding t-SNE was applied, and saliency maps were used to select the spectral regions most influential in classification. The findings reveal that the application of LIBS with deep learning holds a great promise as a fast and non-destructive method in soybean seed variety identification. Further research could advance such portable LIBS devices and further refine ablation techniques for still better detection and identification of this substance.**[22]**
* Soybean is an important crop, produced with the highest protein yield per hectare and an account of over 60% global oilseed production. It is utilized in several industries, such as human food, animal feed, and in the manufacture of biodiesel. Being the world's leading producer, the United States accounts for 29% of the world's soybean production and 19% of soybean meal mainly used as livestock feed. Soybeans, in fact contain most of the essential amino acids necessary for both human and animal nutrition. However, its amino acid profile which characteristically lacks sufficient sulfur-containing amino acids such as methionine and cysteine inhibits its nutritional quality as animal feed especially for monogastric animals.The nutritional improvement of soybean varieties can be taken from improving the balance of amino acids and seed protein content. Historically, selections made for increasing seed oil have also resulted in decreased protein levels, which has proven problematic for optimization toward an optimal oil:protein ratio. Environmental factors such as temperature and water stress can have significant impacts on protein stability across genotypes. To overcome these problems, more than 250 QTLs associated with seed protein content have been identified.Research evidence shows that certain amino acids-the valine, phenylalanine, and threonine, in particular-are involved in maintaining protein stability under different conditions. Targeting these amino acids by plant breeding techniques may improve the quality of soybean in its entirety. Although seed protein content is inversely correlated to oil content, the nutritional value and production can be enhanced.**[23]**

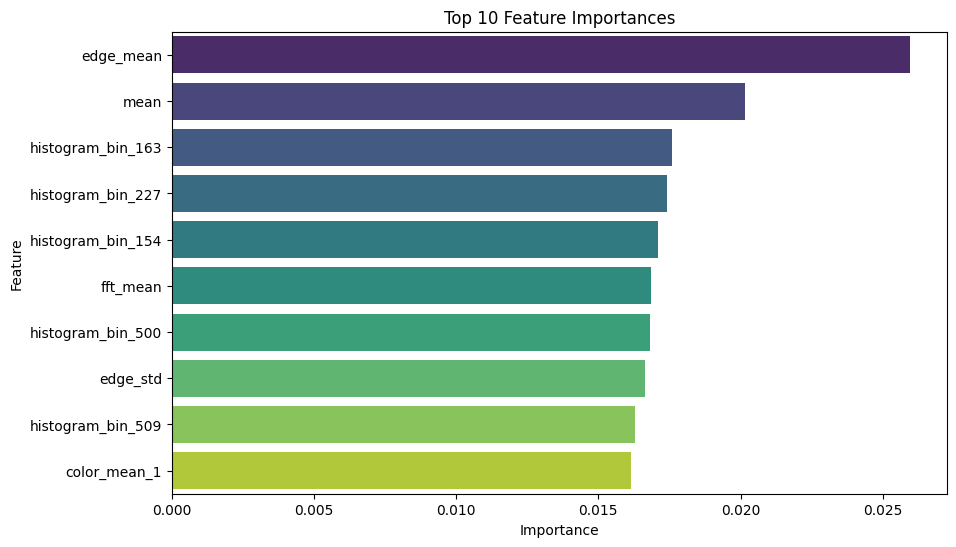
**METHODOLOGY:**

The data set consists of 5500 rows and 536 columns including class label.The attributes are derived from the various images of soybean seed.data set is extracted from examining nearly 5600 images.

|  |  |
| --- | --- |
| **Labels** | **count** |
| **Skin damaged** | **1128** |
| **Intact** | **1201** |
| **Broken** | **1089** |
| **Spotted** | **1058** |
| **Immature** | **1125** |

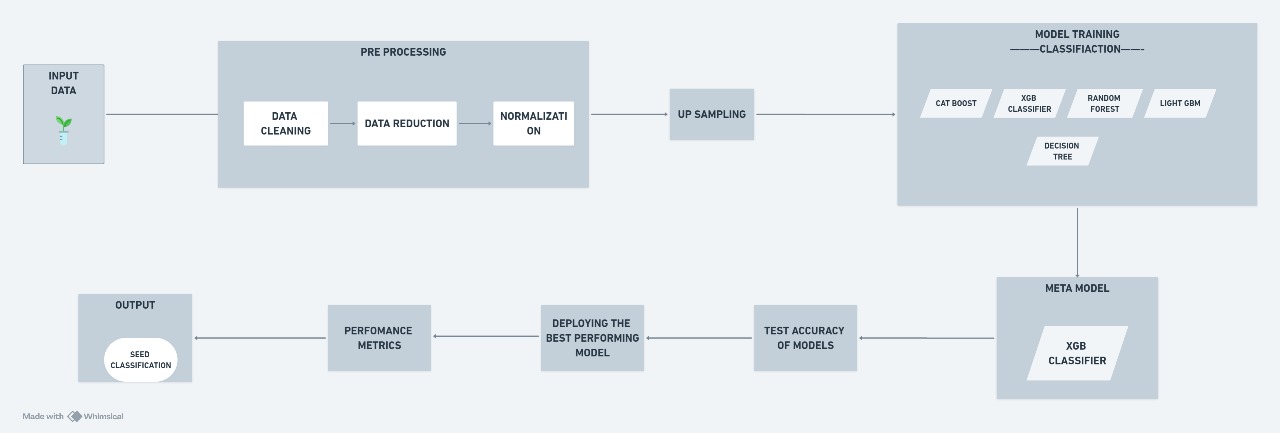
****

****

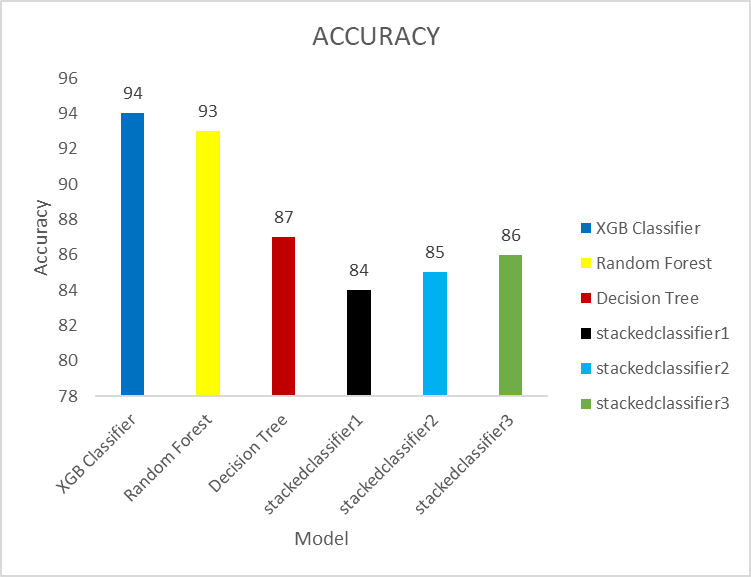
****

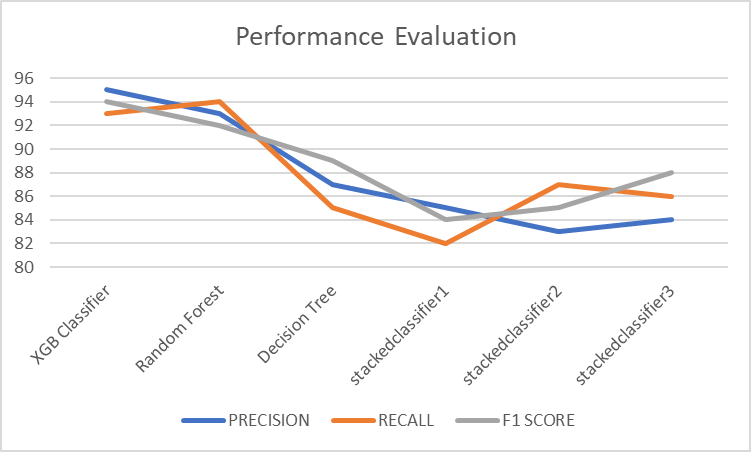
XGB Classifier is an ensemble learning model that improved the accuracy of the

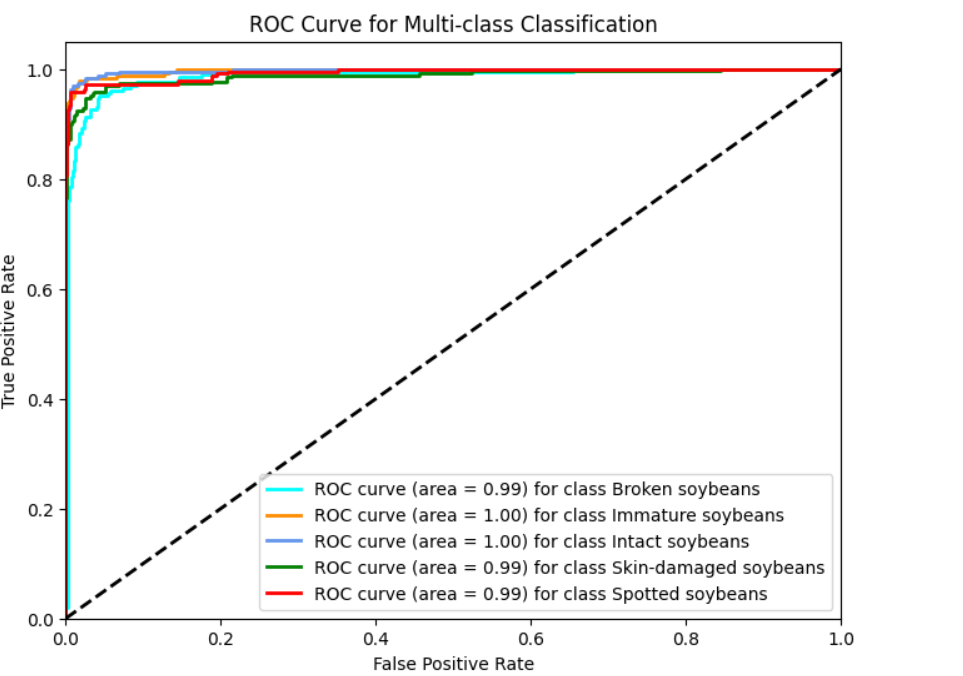
prediction. These base models are initially trained individually on the same dataset. Each model has different accuracy to prediction. Advantages of using this classifier is, the accuracy is improved and also flexibility. We also used up sampling which is a technique used to increase the number of instances in a dataset. The invocation of up sampling technique and utilization of XGB Classifier produced a better accuracy compared to other models and classifiers.

****

**RESULTS AND DISCUSSIONS:**





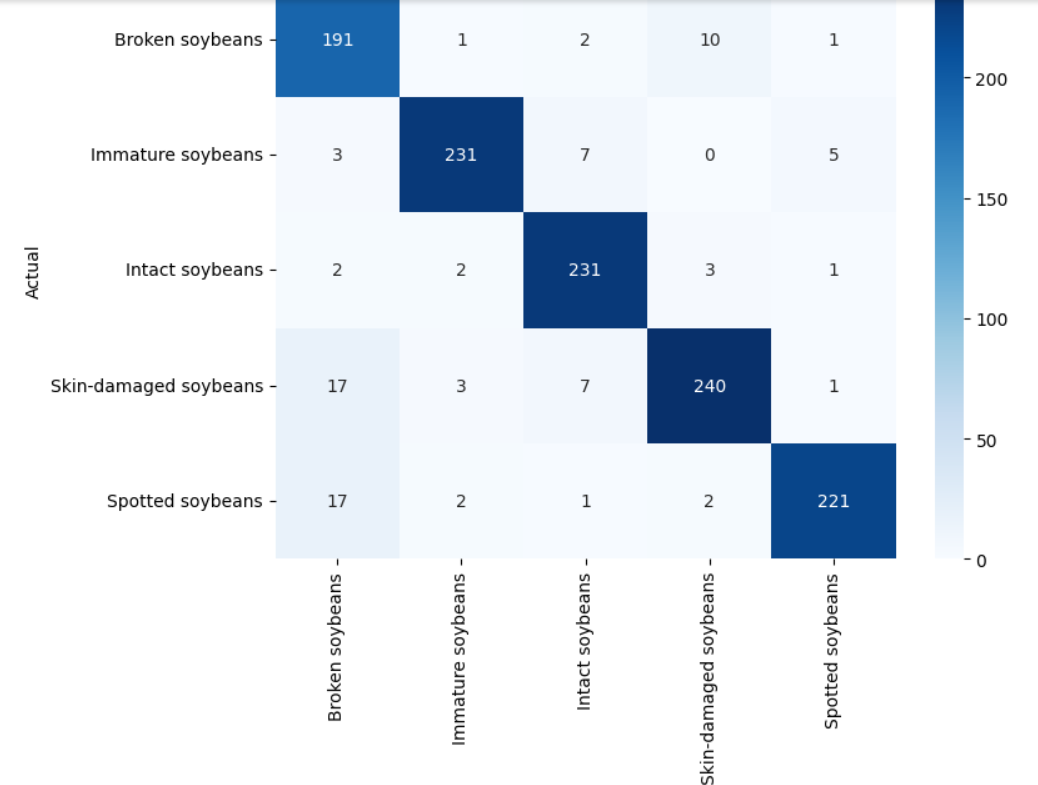


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL | ACCURACY | PRECISION | RECALL | F1 SCORE |
| XGB Classifier | 94 | 95 | 93 | 94 |
| Random Forest | 93 | 93 | 94 | 92 |
| Decision Tree | 87 | 87 | 85 | 89 |
| stackedclassifier1 | 84 | 85 | 82 | 84 |
| stackedclassifier2 | 85 | 83 | 87 | 85 |
| stackedclassifier3 | 86 | 84 | 86 | 88 |

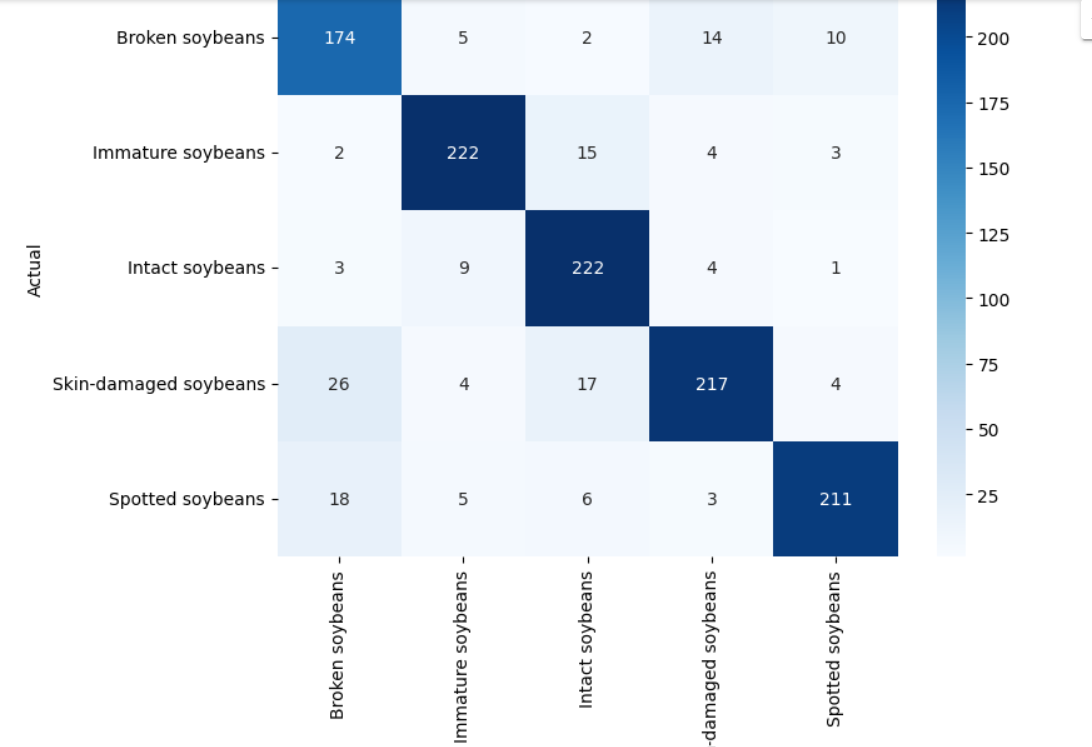
* StackingClassifier1-CatBoost and XGBClassifer
* StackingClassifier2-CatBoost and XGBClassifer and LightGBM
* StackingClassifier3-Random Forest and XGBClassifer and LightGBM

**CONFUSION MATRIX** 

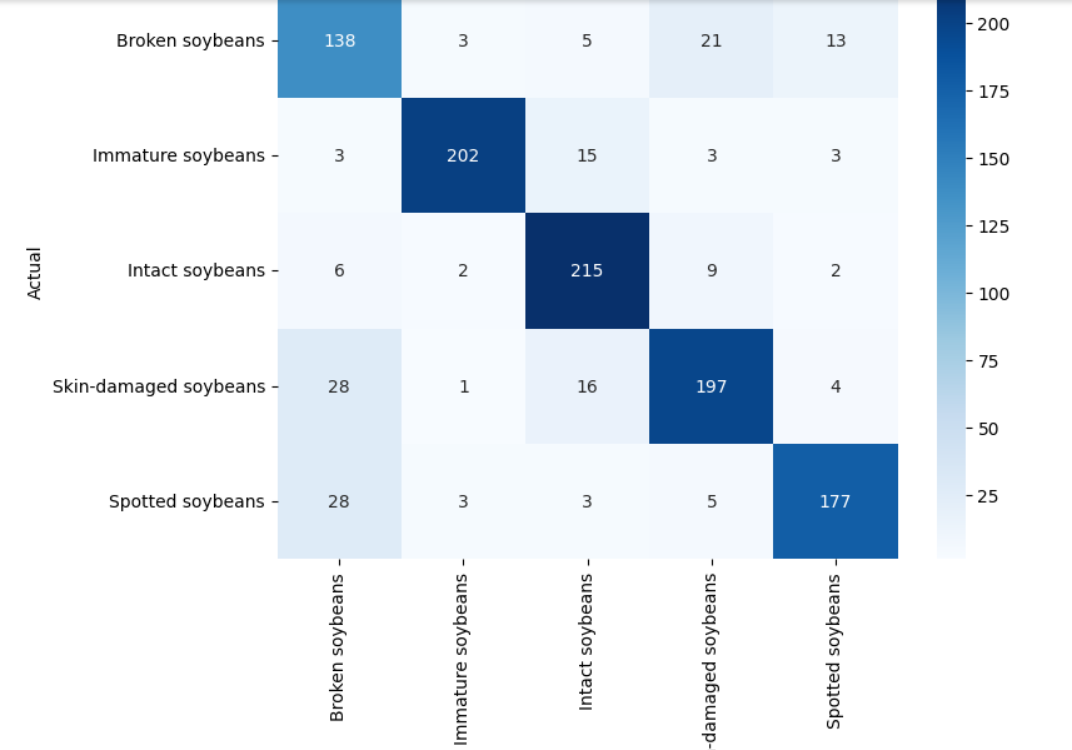
XGBCLASSIFIER



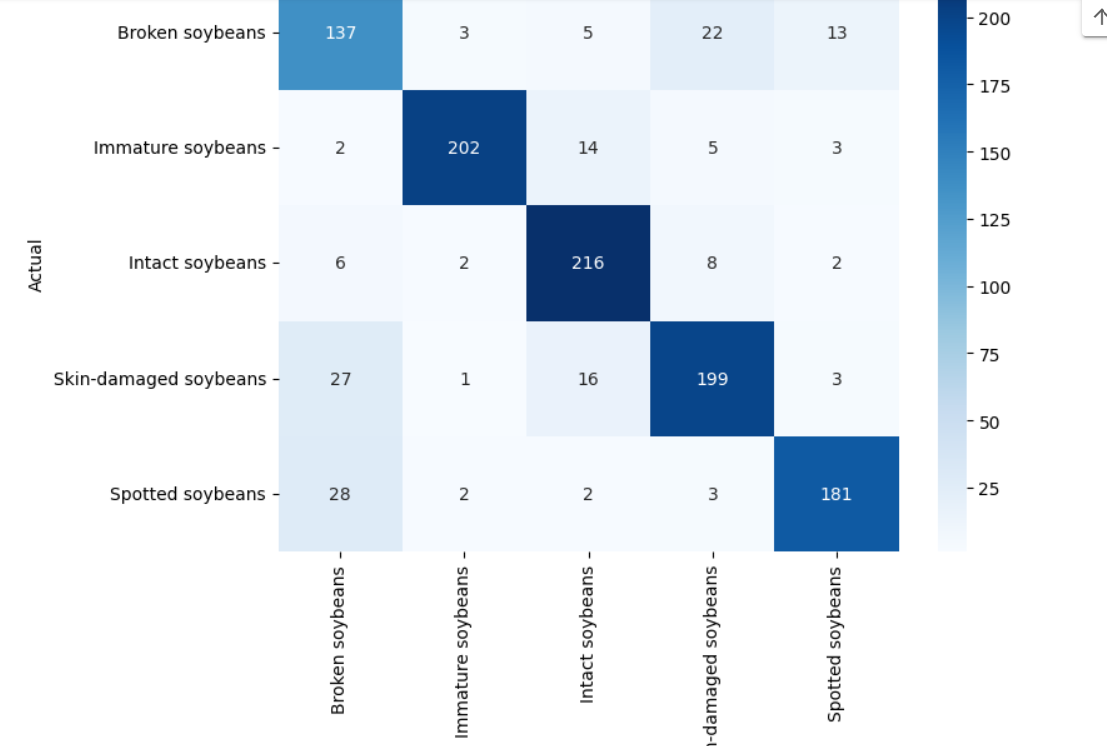
Random Forest



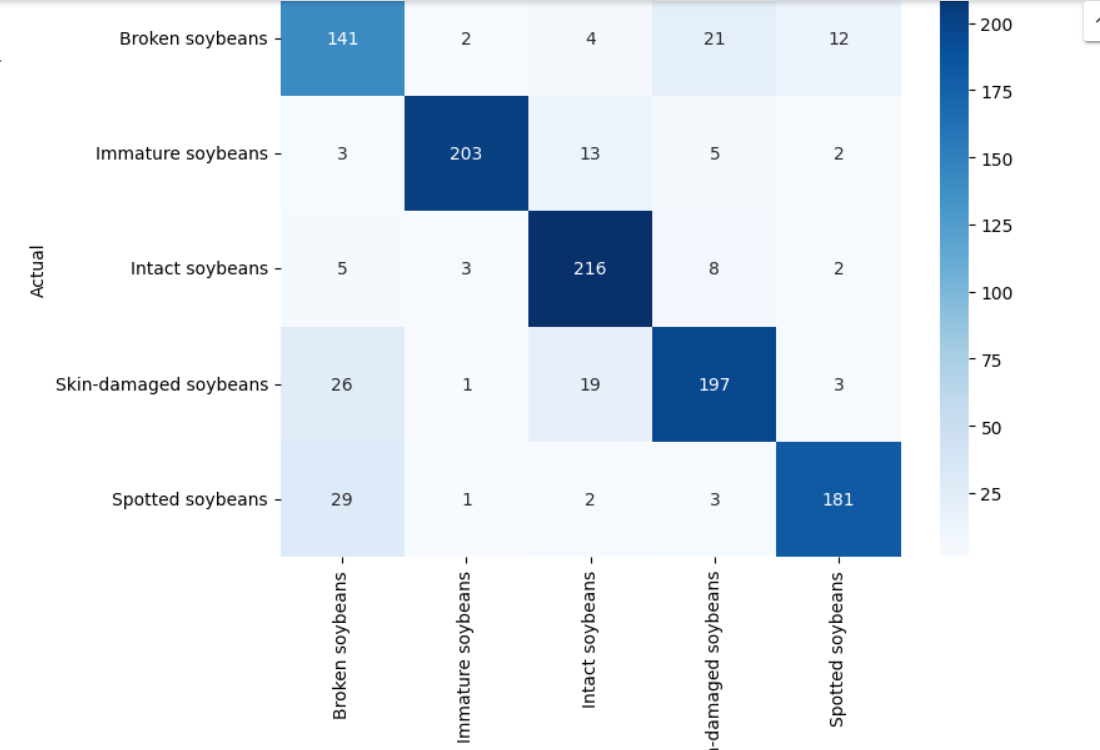
Decision Tree



CatBoost and XGBClassifer



CatBoost and XGBClassifer and LightGBM



Random Forest and XGBClassifer and LightGBM

**REFERENCES:**

* **[1]** Quadras, D. L. de O., Cavalcante, I., Kück, M., Mendes, L. G., & Frazzon, E. M. (2023). Machine Learning Applied to Logistics Decision Making: Improvements to the Soybean Seed Classification Process. Applied Sciences, 13(19), 10904. <https://doi.org/10.3390/app131910904>
* **[2]** Sable, A., Singh, P., Kaur, A., Driss, M., & Boulila, W. (2024). Quantifying Soybean Defects: A Computational Approach to Seed Classification Using Deep LearningTechniques.Agronomy,14(1098). [https://doi.org/10.3390/agronomy14061098&#8203;:contentReference[oaicite:0]{index=0](https://doi.org/10.3390/agronomy14061098&#8203;:contentReference%5Boaicite:0%5D%7Bindex=0)}.
* **[3]** Gadotti, G. I., Ascoli, C. A., Bernardy, R., Monteiro, R. de C. M., & Pinheiro, R. de M. (2022). Machine Learning for Soybean Seeds Lots Classification. Engenharia Agrícola, 42(Special Issue), e20210101. <https://doi.org/10.1590/1809-4430-Eng.Agric.v42nepe20210101/2022>
* **[4]** Ziliang Huang, Rujing Wang, Ying Cao, Shijian Zheng, Yue Teng, Fenmei Wang, Liusan Wang, Jianming Du,Deep learning based soybean seed classification,Computers and Electronics in Agriculture,Volume202,2022,107393,ISSN0168-1699,<https://doi.org/10.1016/j.compag.2022.107393>.
* **[5]** de Medeiros, A.D., Capobiango, N.P., da Silva, J.M. et al. Interactive machine learning for soybean seed and seedling quality classification. Sci Rep 10, 11267 (2020). <https://doi.org/10.1038/s41598-020-6827>
* **[6]** Yang, Y., Liao, J., Li, H., Tan, K., & Zhang, X. (2023). Identification of high-oil content soybean using hyperspectral reflectance and one-dimensional convolutional neural network. Spectroscopy Letters, 56(1), 28-41. <https://doi.org/10.1080/00387010.2022>
* **[7]** Torsoni, G.B., Aparecido, L.E.O., dos Santos, G.M., Chiquito, A.G., Moraes, J.R.S.C., & Rolim, G.S. (2023). Soybean yield prediction by machine learning and climate. Theoretical and Applied Climatology, 151, 1709–1725. [https://doi.org/10.1007/s00704-022-04341-9&#8203;:contentReference[oaicite:0]{index=0](https://doi.org/10.1007/s00704-022-04341-9&#8203;:contentReference%5Boaicite:0%5D%7Bindex=0)}.
* **[8]**Oliveira DC de, Barbosa UC, Bergland ACRO, Resende O, Oliveira DEC de. G-SOJA - WEBSITE WITH PREDICTION ON SOYBEAN CLASSIFICATION USING MACHINE LEARNING. Eng Agríc [Internet]. 2022;42(spe):e20210140. Available from: <https://doi.org/10.1590/1809-4430-Eng.Agric.v42nepe20210140/2022>
* **[9]** Fletcher, R. S. (2016). Using vegetation indices as input into random forest for soybean and weed classification. American Journal of Plant Sciences, 7(15), 2186-2198. [https://doi.org/10.4236/ajps.2016.715193&#8203;:contentReference[oaicite:0]{index=0](https://doi.org/10.4236/ajps.2016.715193&#8203;:contentReference%5Boaicite:0%5D%7Bindex=0)}.
* **[10]** Cetin, N. Machine Learning for Varietal Binary Classification of Soybean (Glycine max (L.) Merrill) Seeds Based on Shape and Size Attributes. Food Anal. Methods 15, 2260–2273 (2022). <https://doi.org/10.1007/s12161-022-02286-3>
* **[11]** Sundaramoorthi, D., Dong, L. Machine learning and optimization based decision-support tool for seed variety selection. Ann Oper Res (2022). <https://doi.org/10.1007/s10479-022-04995-8>
* **[12]** H. Pratap, M. S. Guru Prasad, P. Agarwal, A. Bhardwaj, A. Mehra and S. Mathpati, "SoyNet: Deep Learning Approaches for Automated Soybean Seed Quality Assessment," 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2024, pp. 613-617, doi: 10.1109/InCACCT61598.2024.10551180.
* **[13]** Gulzar, Y., 2024. Enhancing soybean classification with modified inception model: A transfer learning approach. Emirates Journal of Food and Agriculture, 36, pp.1-9.
* **[14]** Wei Lin, Lei Shu, Weibo Zhong, Wei Lu, Daoyi Ma,Yizhen Meng,Online classification of soybean seeds based on deep learning, Engineering Applications of Artificial Intelligence,Volume 123, Part C,2023,106434,ISSN 0952-1976,https://doi.org/10.1016/j.engappai.2023.106434.

(<https://www.sciencedirect.com/science/article/pii/S0952197623006188>)

* **[15]** Melero, F. C., Masetto, T. E., Rocha, L. G., Silva, D. M., Crispim, B. A., & Barufatti, A. (2024). DNA degradation is involved with low physiological potential of soybean. *Sci. Agric.*, 81, e20230250. [**https://doi.org/10.1590/1678-992X-2023-0250**](https://doi.org/10.1590/1678-992X-2023-0250)**.**
* **[16]** Park, Y.-H., & Jun, T.-H. (2022). Classification of Soybean [Glycine max (L.) Merr.] Seed Based on Deep Learning Using the YOLOv5 Model. Plant Breed. Biotech., 10(1), 75-80. [https://doi.org/10.9787/PBB.2022.10.1.75&#8203;:contentReference[oaicite:0]{index=0](https://doi.org/10.9787/PBB.2022.10.1.75&#8203;:contentReference%5Boaicite:0%5D%7Bindex=0)}.
* **[17]** Yan, H., Tian, D., Zhang, Q., Wen, J., Wang, Z.-Y., & Chai, M. (2024). GmHs1-1 and GmqHS1 Simultaneously Contribute to the Domestication of Soybean Hard-Seededness. Plants, 13(15), 2061. <https://doi.org/10.3390/plants13152061>
* **[18**] Wang, M., Ding, X., Zeng, Y., Xie, G., Yu, J., Jin, M., Liu, L., Li, P., Zhao, N., Dong, Q., Liu, B., & Xu, C. (2024). Identification of multiple genetic loci and candidate genes determining seed size and weight in soybean. Agronomy, 14(9), 1957. <https://doi.org/10.3390/agronomy14091957>
* **[19]** Chen, X., He, W., Ye, Z., Gai, J., Lu, W., & Xing, G. (2024). Soybean seed pest damage detection method based on spatial frequency domain imaging combined with RL-SVM. Plant Methods, 20, 130. [https://doi.org/10.1186/s13007-024-01257-5&#8203;:contentReference[oaicite:0](https://doi.org/10.1186/s13007-024-01257-5&#8203;:contentReference%5Boaicite:0)]{

index=0}.

* **[20]** Liang, S., Duan, Z., He, X., Yang, X., Yuan, Y., Liang, Q., Pan, Y., Zhou, G., Zhang, M., Liu, S., & Tian, Z. (2024). Natural variation in GmSW17 controls seed size in soybean. *Nature Communications, 15*(1), 7417. [**https://doi.org/10.1038/s41467-024-51798-5**](https://doi.org/10.1038/s41467-024-51798-5)
* **[21]** Zhang, L., Cao, Z., Gao, Y., Huang, W., Si, Z., Guo, Y., Wang, H., & Wang, X. (2024). Soybean Yield Simulation and Sustainability Assessment Based on the DSSAT-CROPGRO-Soybean Model. Plants, 13(17), 2525. <https://doi.org/10.3390/plants13172525>
* **[22]** Li X, He Z, Liu F, Chen R. Fast Identification of Soybean Seed Varieties Using Laser-Induced Breakdown Spectroscopy Combined With Convolutional Neural Network. Front Plant Sci. 2021 Oct 6;12:714557. doi: 10.3389/fpls.2021.714557. PMID: 34691095; PMCID: PMC8527016.
* **[23]** Wu, C., Acuña, A., Florez-Palacios, L., Harrison, D., Rogers, D., Mozzoni, L., Mian, R., & Canella Vieira, C. (2024). Across-environment seed protein stability and genetic architecture of seed components in soybean. Scientific Reports, [https://doi.org/10.1038/s41598-024-67035-4&#8203;:contentReference[oaicite:0]{index=0](https://doi.org/10.1038/s41598-024-67035-4&#8203;:contentReference%5Boaicite:0%5D%7Bindex=0)}.